National Aeronautics and Space Administration

Performance Evaluation of Pleiades Broadwell Nodes Using NASA Applications

May 5, 2016

NASA Advanced Supercomputing Division

Outline

- Architectural features of Pleiades: Five generations of nodes
- Why core frequency is decreasing and hardware parallelism is increasing?
- Hardware features:
  - ✓ Turbo-Boost
  - ✓ Hyper-Threading (HT)
  - ✓ Wider SIMDs (SSE4 vs. AVX vs. AVX2): Number of Flops per cycle
  - ✓ Clock frequency, Turbo frequency and AVX frequency
- Applications:
  - ✓ FUN3D
  - ✓ USM3D
  - ✓ Overflow
  - ✓ Cart3D
  - ✓ MITgcm
- Results:
  - ✓ Memory bandwidth per core
  - ✓ Floating-point efficiency
  - ✓ Turbo-Boost:
  - ✓ Hyper-Threading (HT)
  - ✓ AVX vs. AVX2
  - ✓ FUN3D, USM3D, Overflow, Cart3D and MITgcm
- Modeling: Upper bound efficiency of BLAS 1 (AXPY and DOT).
- Conclusions

Challenges to Application Software - Parallelism

	Harpertown (HPT) 11/2007	Nehalem (NHM) 03/2009	Westmere (WES) 04/2011	Sandy Bridge (SNB) 03/2012	Ivy Bridge (IVB) 09/2013	Haswell (HAS) 09/2014	Broadwell (BDW) 03/2016
Core(s)	4	4	6	8	10	12	14
Threads	4/8	8	12	16	20	24	28
SIMD Width	128	128	128	256	256	2x 256 (FMA3)	2x 256 (FMA3)
CPU-Clock (GHz)	3.0	2.93	2.93	2.6	2.8	2.5	2.4

- Number of cores **increased by 40%** from 4 in Harpertown to 14 in Broadwell.
- Clock speeds **decreased by 20%** 3.0 GHz of Harpertown to 2.4 GHz in Broadwell.
- Number of cores **increased by 40%** from 4 in Harpertown to 14 in Broadwell.
- Number of threads **increased by 600%** from 4 in Harpertown to 28 in Broadwell

More cores → **More threads** → **Wider vectors**

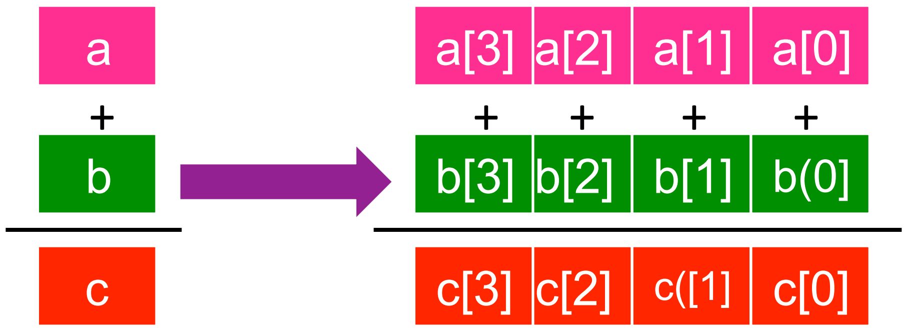
Why CPU clock speed is decreasing and parallelism is increasing?

Clock, Turbo and AVX speeds

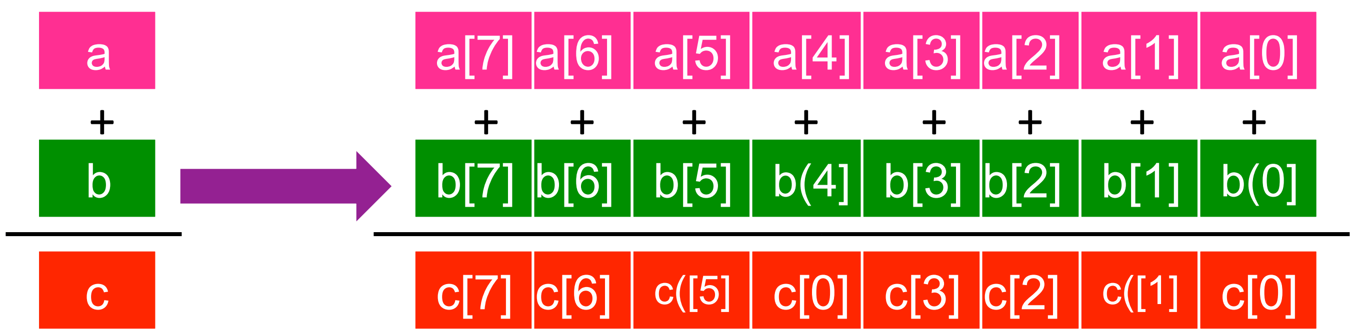
	Haswell	Broadwell
Clock speed	2.5 GHz	2.4 GHz
Turbo clock speed	3.3 GHz	3.2 GHz
AVX2 clock speed	2.1 GHz	2.0 GHz
Thermal Design Power (TDP)	120 W	120 W

SSE vs. AVX vs. AVX2

- Streaming SIMD Extensions (SSE):
  - 4 floating point, single precision (32-bit) elements.
  - 2 floating point, double precision (64-bit) elements.
  - SSE instructions operate on all data items in parallel.



- Advanced Vector Extensions (AVX):
  - 8 floating point, single precision (32-bit) elements.
  - 4 floating point, double precision (64-bit) elements.
  - AVX instructions operate on all data items in parallel.



- Advanced Vector Extensions 2 (AVX2): 8 floating point, double precision (64-bit) elements

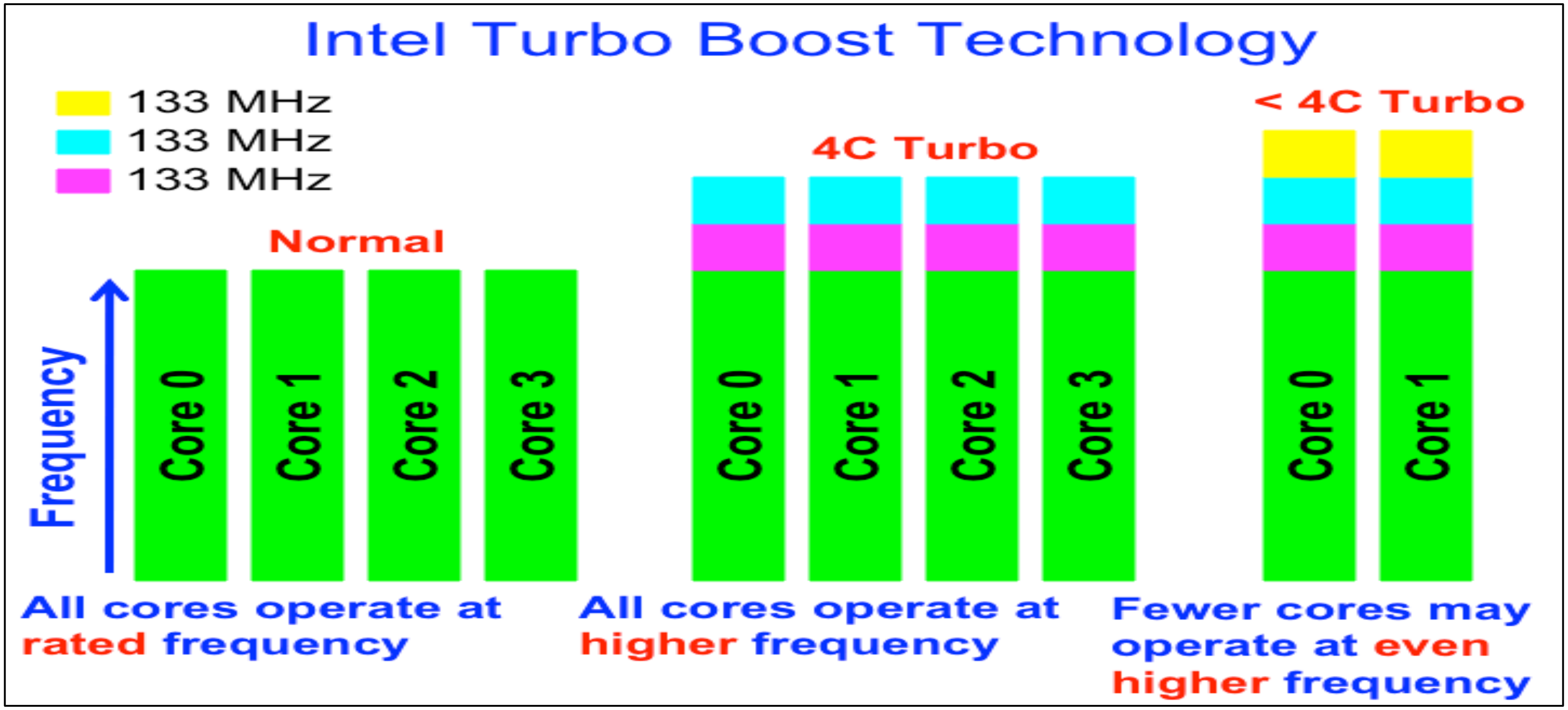
Peak Performance – Per Core

- Most of the recent computers have FMA (Fused multiply add), i.e. x --> x + y * z. It is available on Haswell and Broadwell and known as FMA3 where 3 stands for three operands.

  - All Intel Xeon earlier models have SSE2
    - ✓ 2 flops/cycle in DP.
  - Intel Xeon Nehalem (2009) & Westmere (2012) have SSE3
    - ✓ 4 flops/cycle in DP.
  - Intel Xeon Sandy Bridge (2011) & Ivy Bridge (2012) have AVX
    - ✓ 8 flops/s cycle in DP.
  - Intel Xeon Haswell (2014) & Broadwell (2016) have AVX2
    - ✓ 16 flops/cycle in DP.

- $$\text{FLOPS} = \text{cores} \times \text{clock} \times \frac{\text{FLOPS}}{\text{Cycles}}$$

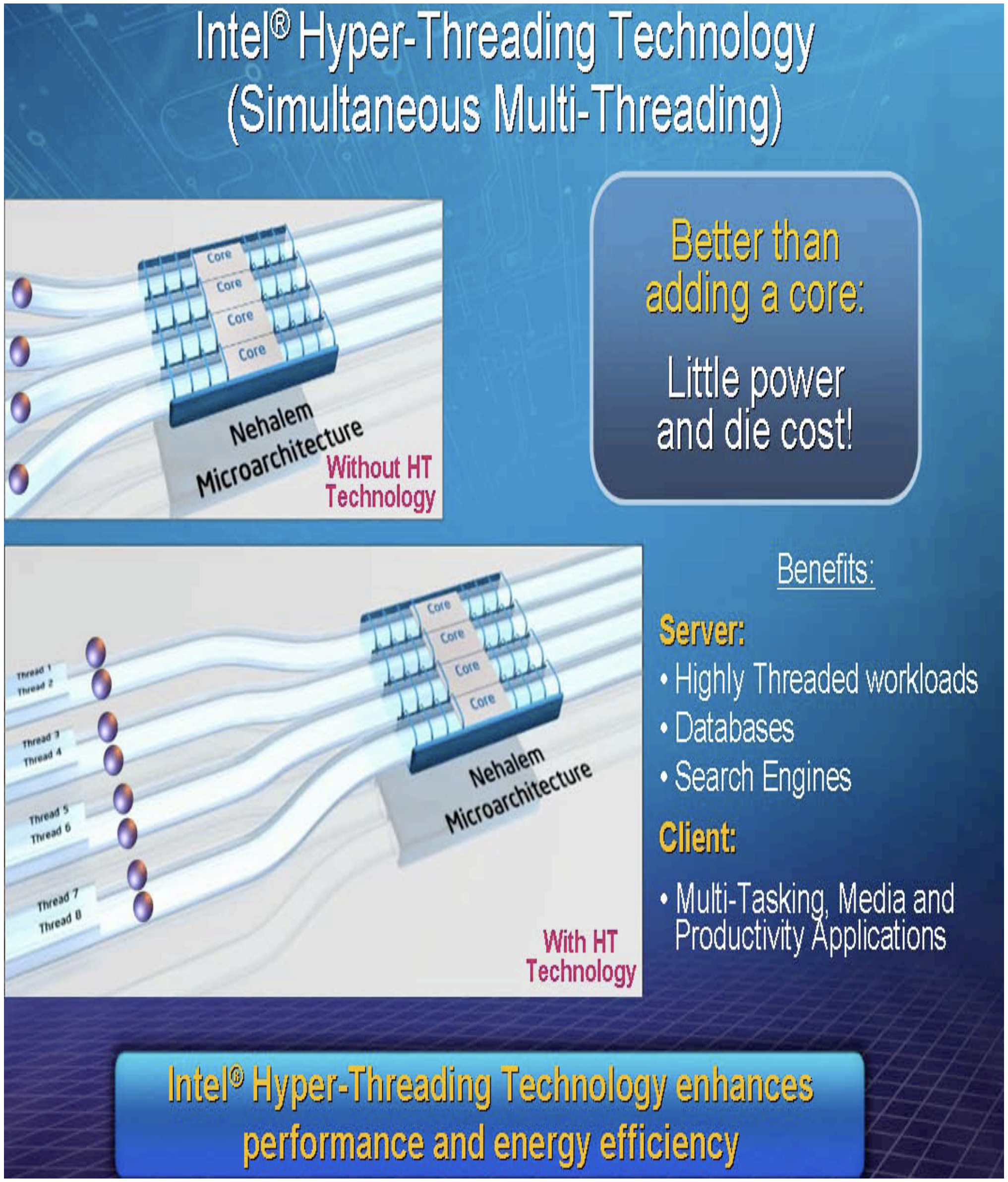
Turbo Boost 1.0 vs. 2.0



- Turbo Boost 1.0: Dynamically increased the frequency of active cores based on temperature, current power consumption, and operating system states. It did not, however, exceed programmed power limits.
- Turbo Boost 2.0: Allows the processor to exceed its power ceiling in a burst, until it reaches its thermal limit, at which point it reduces power to conform to those same programmed limits.

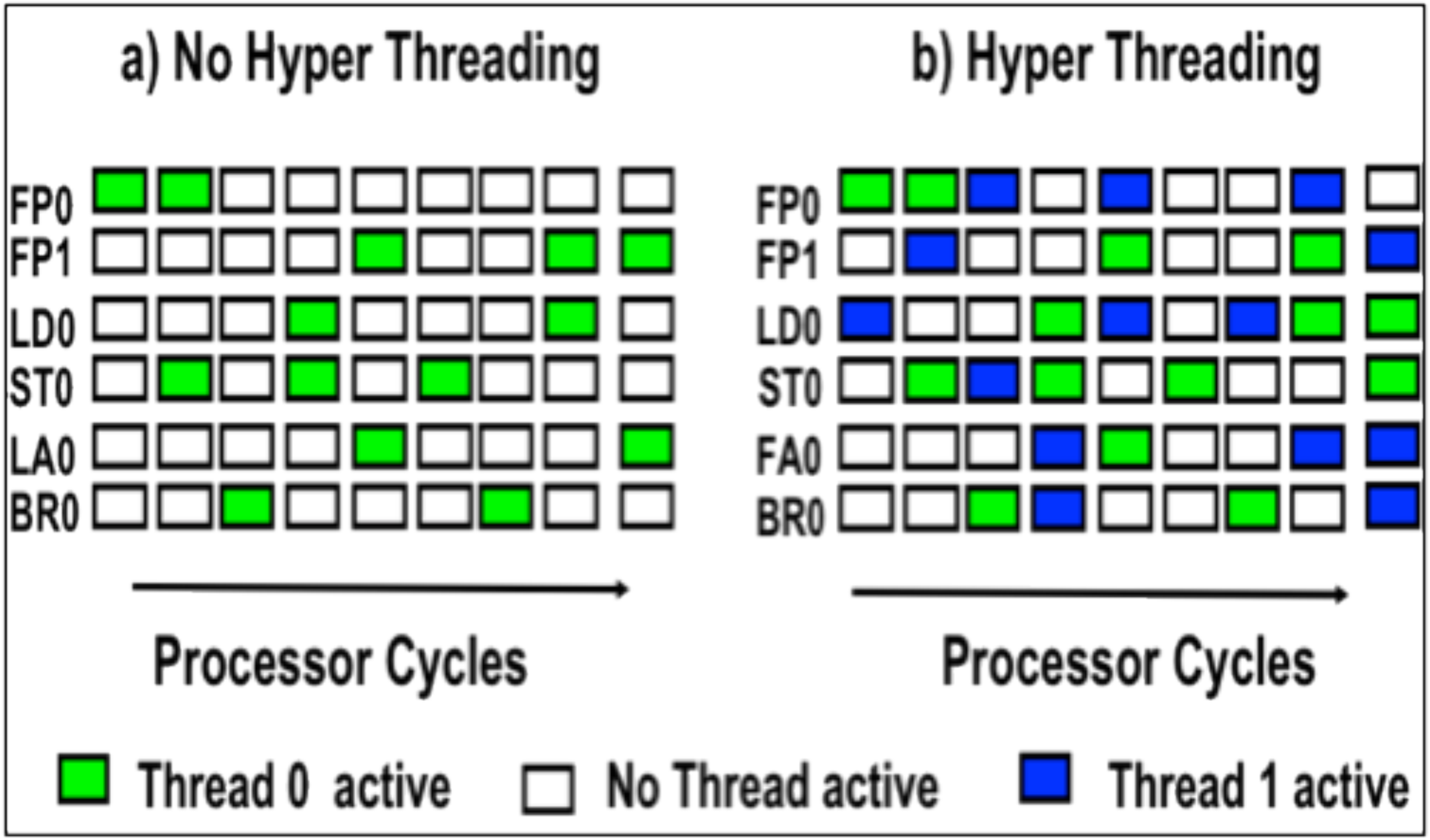
Intel Hyper-Threading Technology

- Also known as SMT
  - Runs 2 threads at the same time per core
- Takes advantage of 4-wide execution engine
  - Keep it fed with multiple threads
  - Hide latency of a single thread
- Power efficient performance
  - Very low die cost
  - Can provide significant performance benefit depending on application
  - Much more efficient than adding an entire core



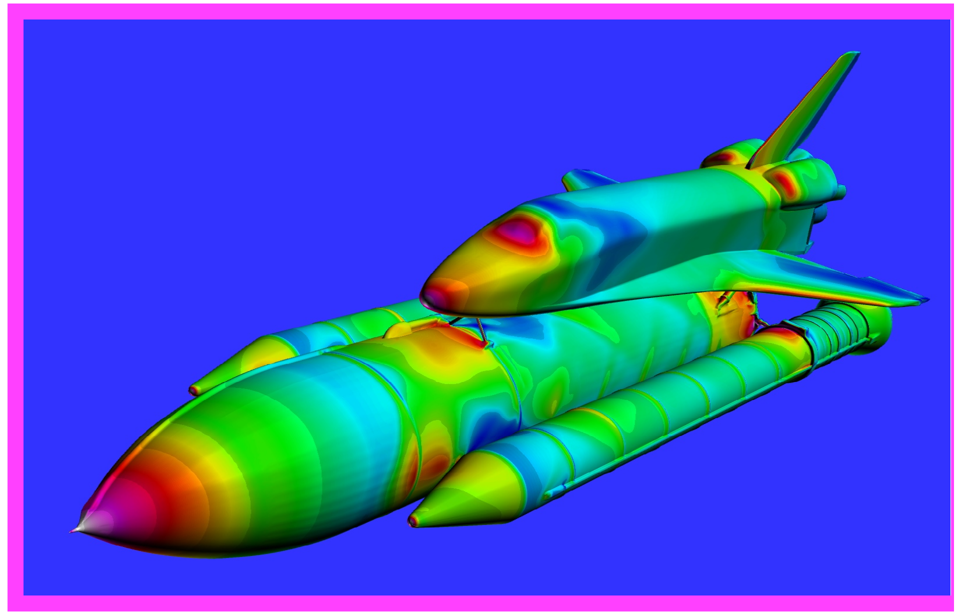
Intel Hyper-Threading (HT)

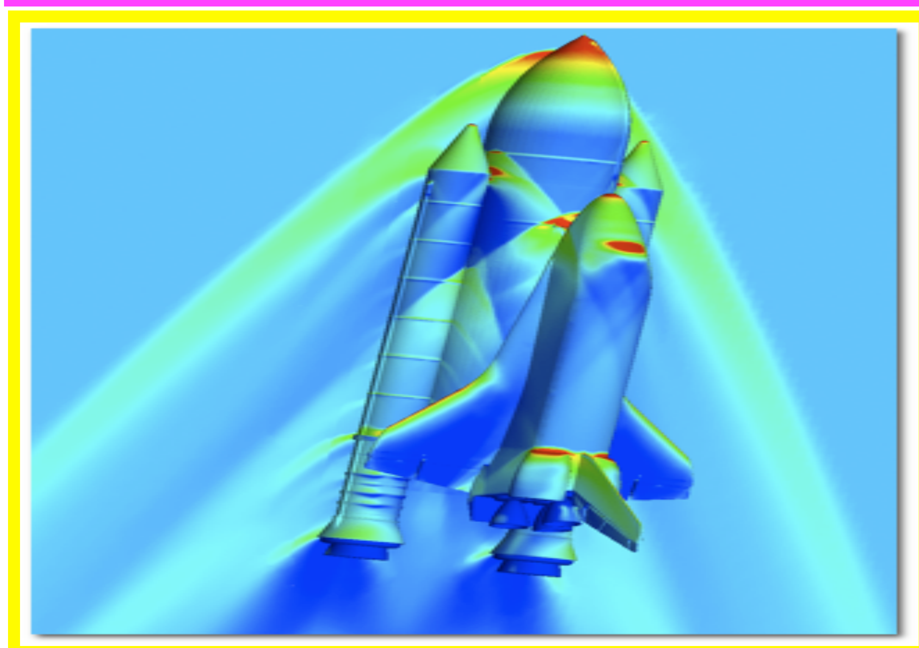
- In HT, operating system (OS) sees two threads on each core.
- Efficient utilization of processor resources.
- Threads share caches and memory bandwidth



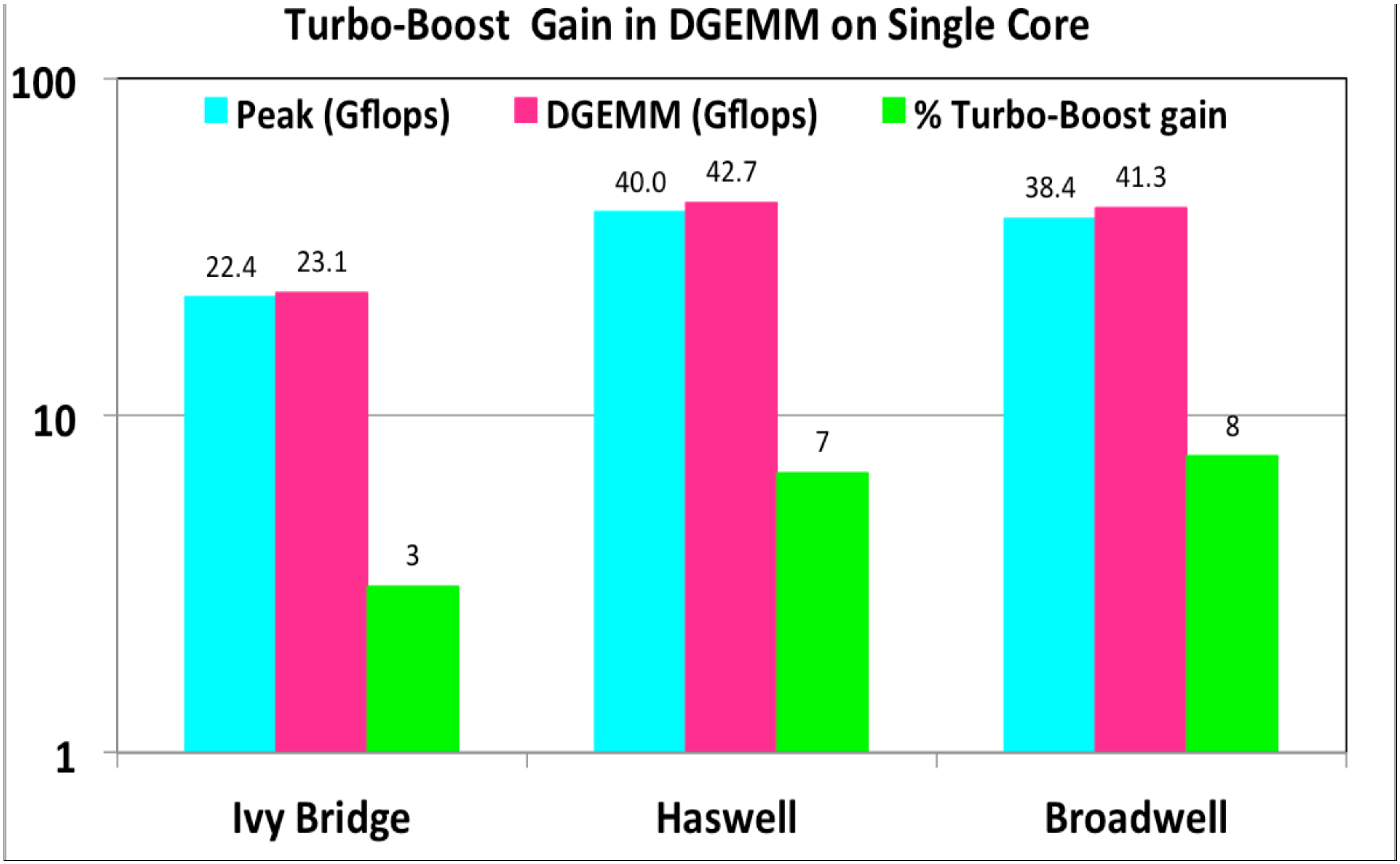
Applications

- **OVERFLOW-2** is a general-purpose Navier-Stokes solver for CFD problems. Data set used is nasrotor grid, 36 and 90 millions grid points.
- **USM3D** is a 3-D unstructured tetrahedral, cell-centered, finite volume Euler and Navier-Stokes flow solver. Data set used is 3D wing configurations, 10 millions and 102 millions cells.
- **FUN3D:** UN3D is an unstructured-grid computational fluid dynamics suite used to tackle NASA's most complex aerodynamics problems. Data set used is 3D wing configurations, 100 millions tetrahedral nodes
- **CART3D** is a high fidelity, inviscid CFD application that solves the Euler equations of fluid dynamics. 24 millions grid points.
- **MITgcm** is a global ocean simulation model for solving the fluid equations of motion. 50M grid points.

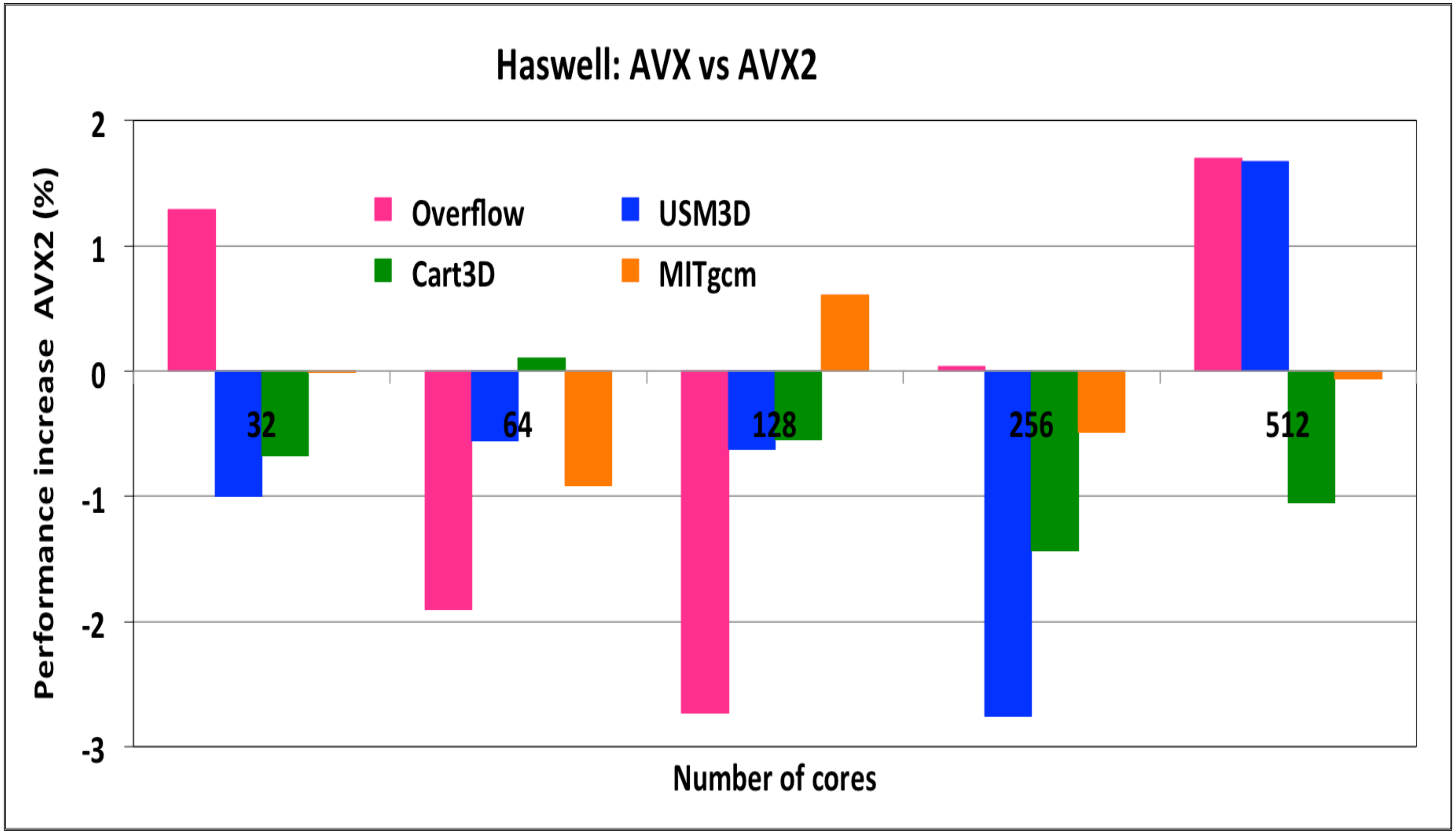




Turbo-Boost Gain on Single Core

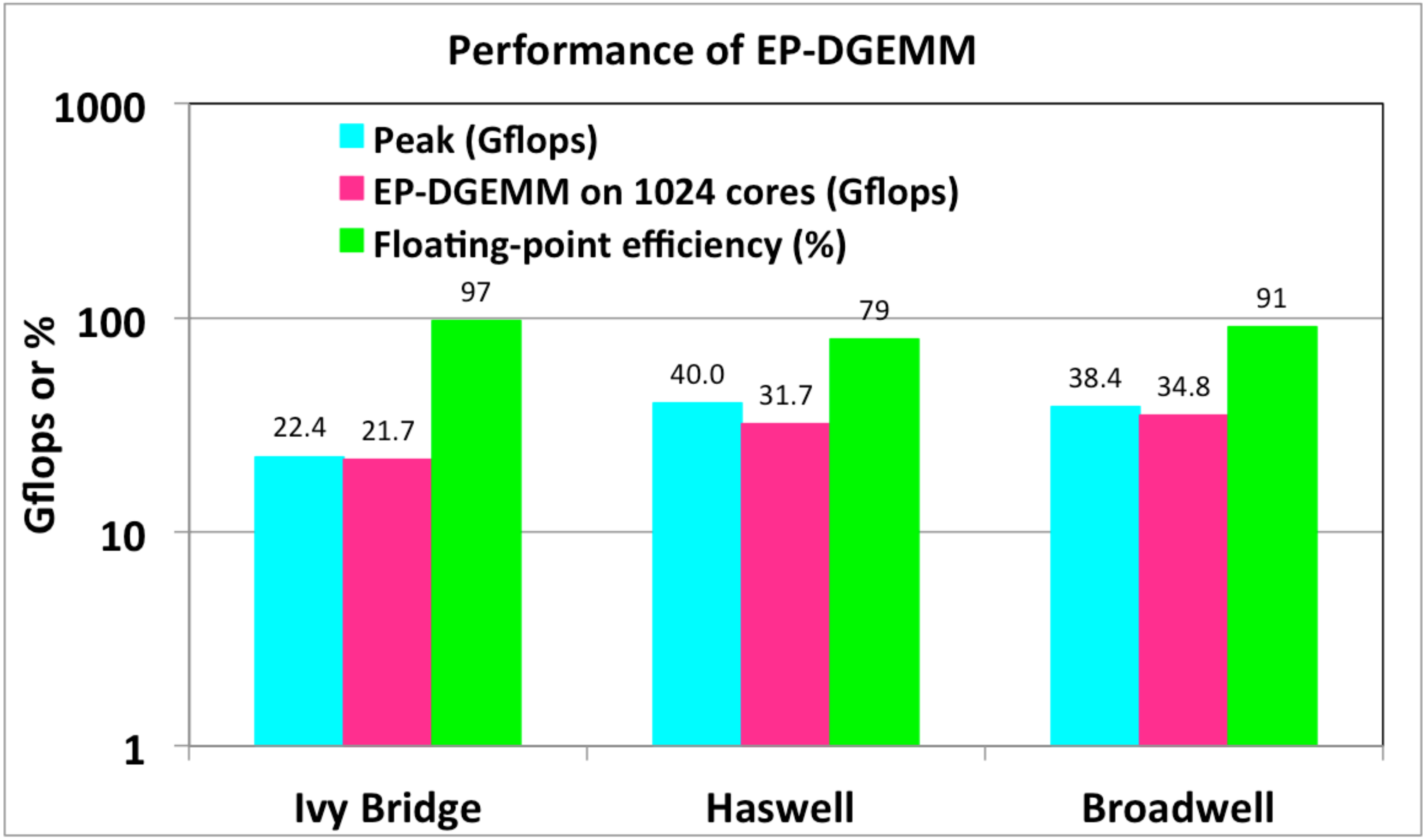


Haswell: AVX vs. AVX2



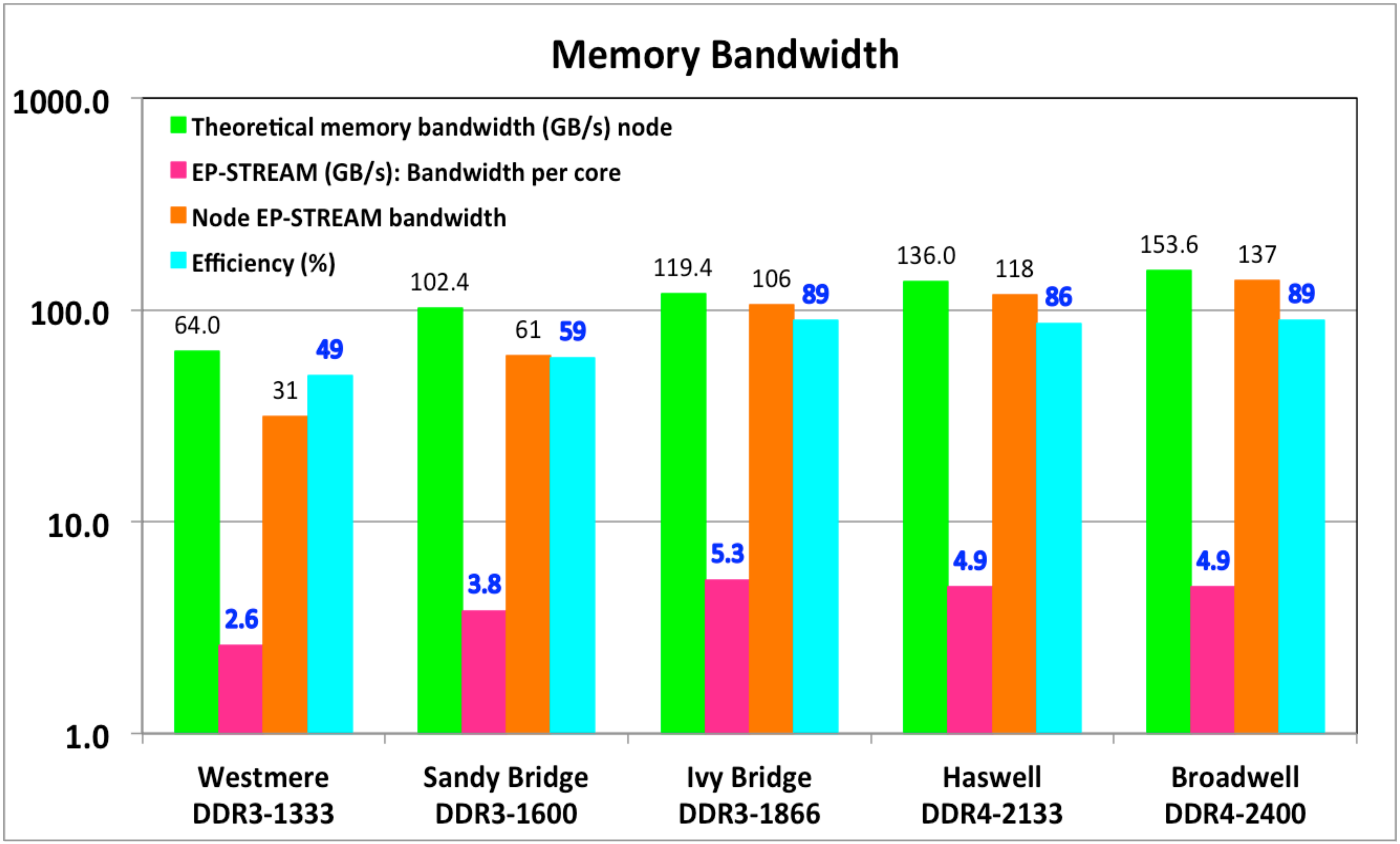
The advantage of AVX2 over AVX instructions is insignificant—ranging from –3 to +2% on the four applications.

Performance of EP-DGEMM (1024 cores)

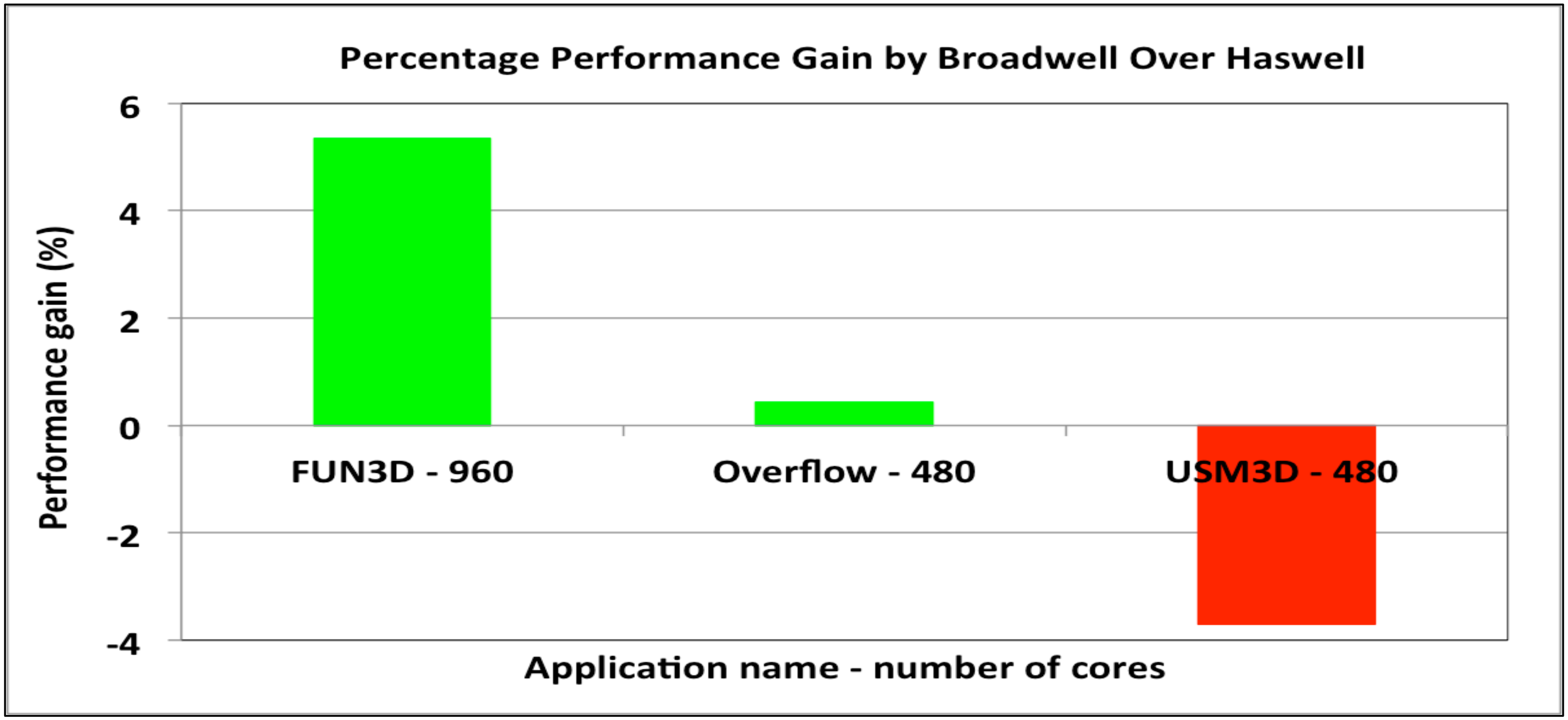


- Broadwell has higher floating-point efficiency: Haswell: 70%; 91% Broadwell.
- EP-DGEMM: Higher performance on Broadwell: Haswell: 31.7 Gflops; 34.8 Gflops in spite of lower peak performance on Broadwell: Haswell: 40 Gflops; 38.4 Gflops.
- The probable reason is due to lower degradation of AVX frequency when AVX2 instruction is issued.

Memory Bandwidth

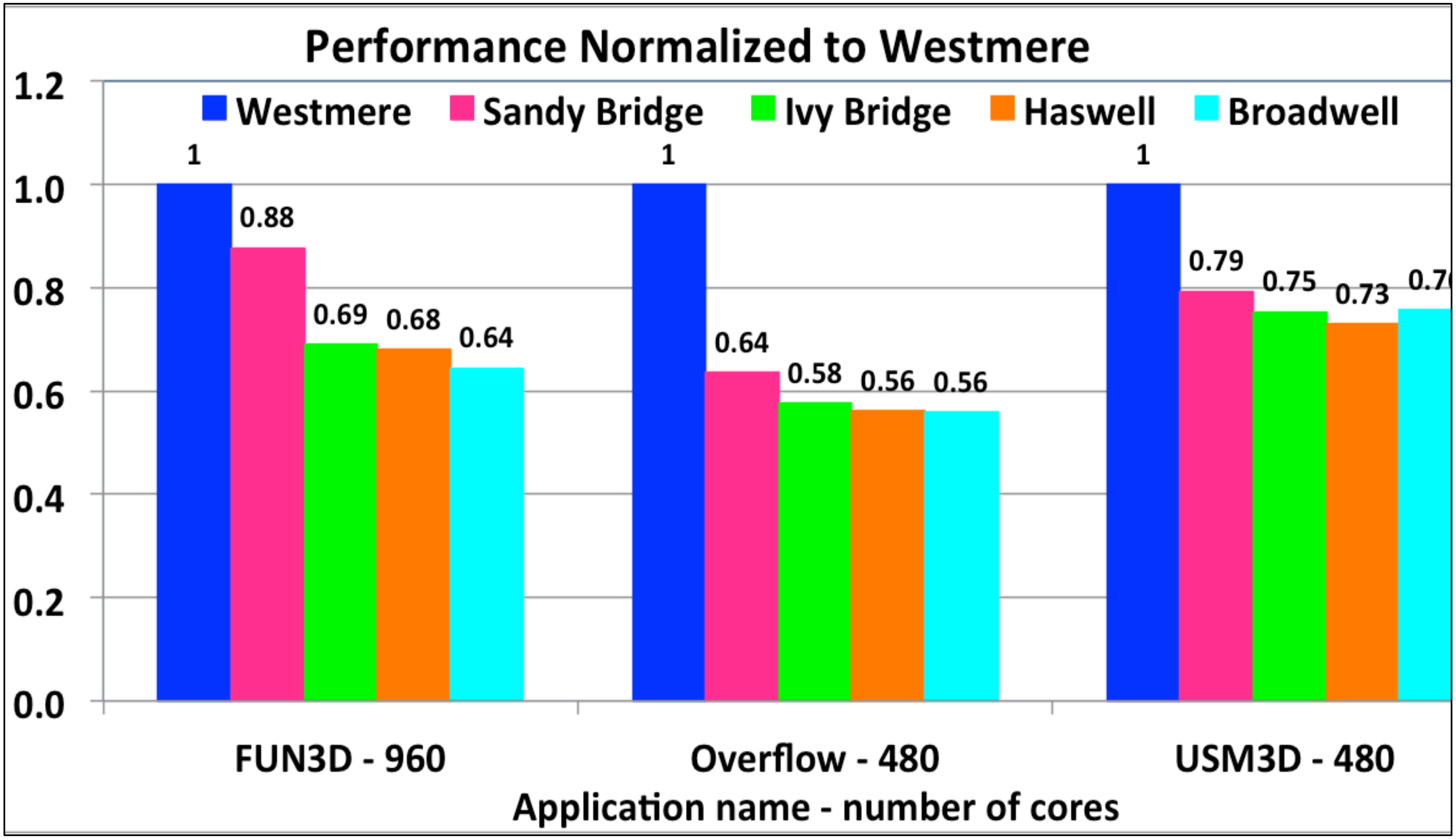


Performance Gain by Broadwell Over Haswell



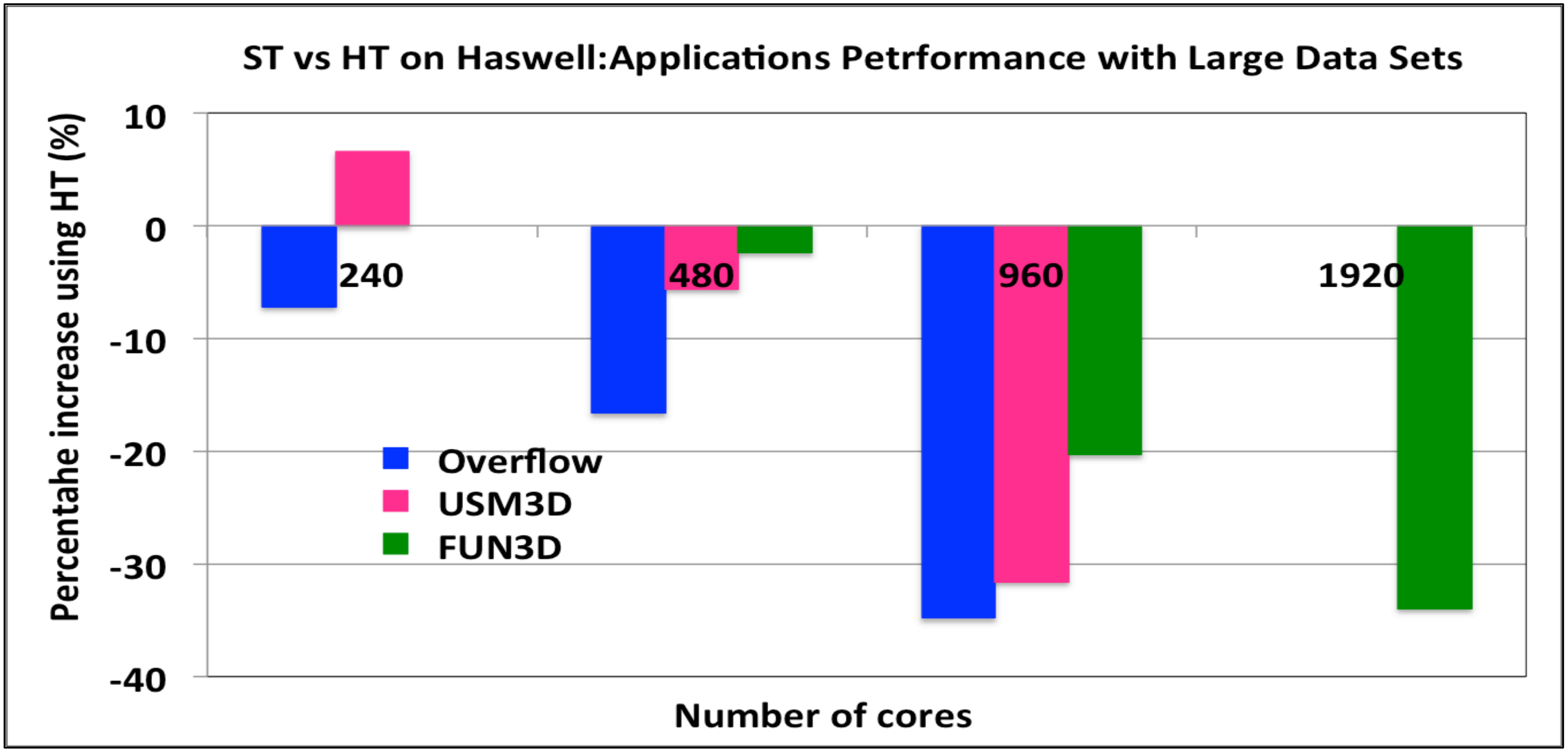
- 480 cores: Performance of USM3D is 4% **lower** on Broadwell than on Haswell.
- 480 cores: Performance of Overflow is almost **same** on Broadwell and Haswell.
- 960 cores: Performance of FUN3D is 5% **higher** on Broadwell than on Haswell
- For large number of cores with MPI collectives, performance on Broadwell is much higher than Haswell due to better node density.
  - ✓ 960 cores: 40 Haswell nodes and 35 Broadwell nodes);
  - ✓Fewer nodes mean less inter-node (node to node), and more intra-node (CPU to CPU) and intra-CPU (core to core) communication.

Application Performance Normalized to Westmere



Lower is better

ST vs. HT on Haswell



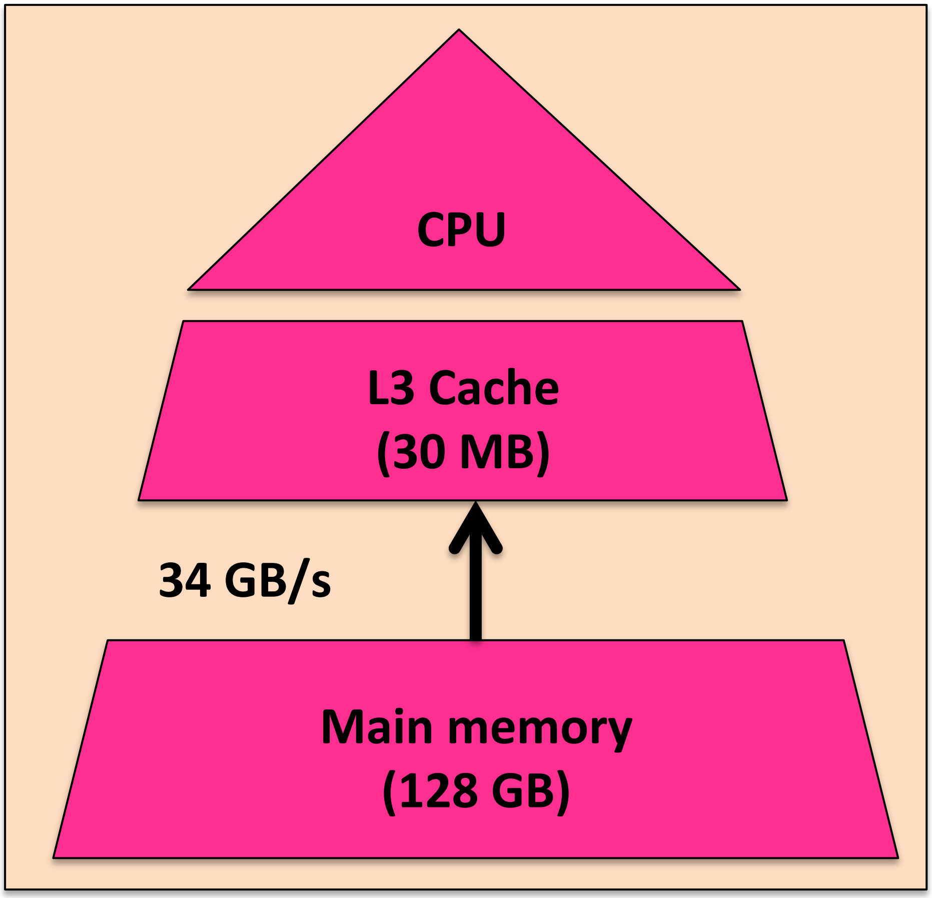
- Threads share L1, L2, and L3 caches.
- Size of L1, L2 and L3 cache per core is same on Pleiades nodes.
- Threads share memory subsystem bandwidth.
- Threads put more pressure on the Host Bus Adapter (HBA) of a node resulting in a bottleneck at HBA.

Memory Transfer

(Its All About Data Movement)

One level of memory

Pleiades Haswell
Clock 2.5 GHz
Turbo Boost 3.3 GHz
16 flops/cycle x
3.3 x 10^9 cycles/sec =
52.8 Gflop/s

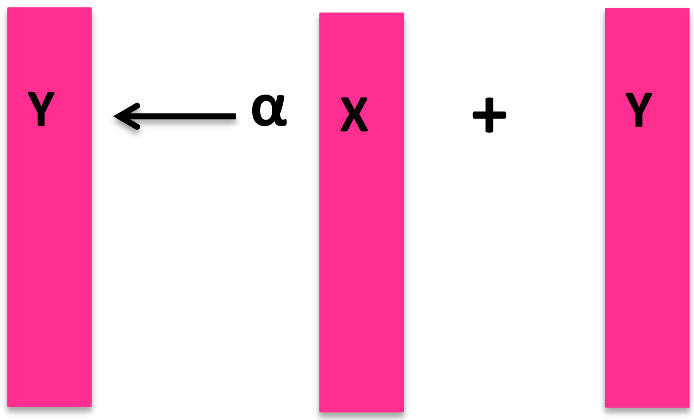


Omitting latency here

The model is simplified (see next slide) but it provides an upper bound on performance as well, i.e. we will never go faster than what the model predicts.

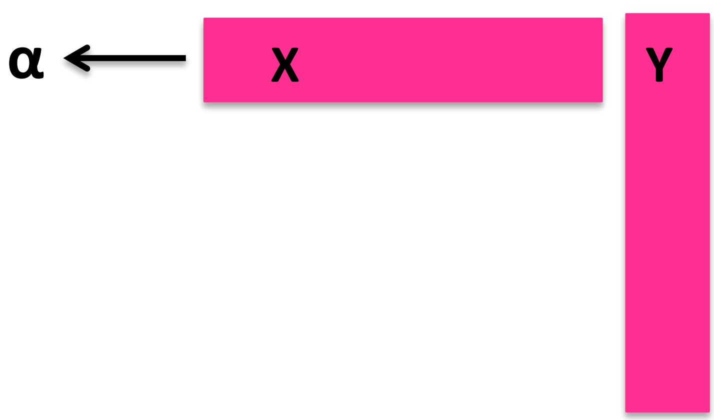
BLAS 1: AXPY and DOT

- **AXPY:**



for (j = 0; j < n; j++)	**n MUL**
y[i] += a * x[i];	**n ADD**
	2n FLOP
(without increment)	**n FMA**

- **DOT:**



alpha = 0e+00	**n MUL**
for (j = 0; j < n; j++)	**n ADD**
alpha += x[i] * y[i];	**2n FLOP**
(without increment)	**n FMA**

Vector Operations

- Take two double precision vectors x and y of size n = 1,875,000

- Data size:
  - (1,875,000 double * (8 Bytes / double) = 15 MB per vector
  - Two vectors fit in cache (30 MB) OK
- Time to move the two vectors from memory to cache
  (30 MB) / (34 GB/s) = **0.88 ms**
- Time to perform computation of DOT
- (2n flop) / (52.8 Gflop/s) **= 0.07 ms**

Vector Operations

- Total_time ≥ max (time_communication, time_compute)

  = max (0.88ms, 0.07ms)

  = 0.88 ms

  Performance = (2 x 1,875,000 flops)/0.88 ms = 4.26 Gflop/s

  Performance for DOT ≤ 4.26 Gflop/s

  Peak is 52.8 Gflop/s

  Efficiency = (4.26 Gflop/s / 52.8 Gflop/s) x 100 = 7.6%

  Efficiency of DOT on Haswell = 8%

Conclusions

- **Architecture**:
  - Architecturally both Haswell and Broadwell are same except for
    - DDR4 speed (2133 vs.2400 MHz). Sustained memory bandwidth is 5.0 vs. 4.9 GB/s.
    - Node density (24 vs. 28 cores).
    - CPU clock (2.5 GHz vs. 2.4 GHz). Peak performance per core is 40.0 vs. 38.4 Gflop/s.
    - Memory per core (5.3 GB vs. 4.6 GB)
    - Better power management on Broadwell.. AVX frequency is less on Broadwell than on Haswell degradation while using AVX2 instruction
- **Performance:**
  - 480 cores: Performance of USM3D is 4% **lower** on Broadwell than on Haswell.
  - 480 cores: Performance of Overflow is **same** on Broadwell and Haswell.
  - 960 cores: Performance of FUN3D is 5% **higher** on Broadwell than on Haswell
  - For large number of cores with MPI collectives, performance on Broadwell is much higher than Haswell due to better node density.
    - ✓ 960 cores: 40 Haswell nodes and 35 Broadwell nodes); 8192 cores: 342 nodes Haswell and 293 nodes.
    - ✓ Fewer nodes mean less inter-node (node to node), and more intra-node (CPU to CPU) and intra-CPU (core to core) communication.
  - Floating-point efficiency of Broadwell is higher than that of Haswell due to lower degradation of AVX frequency.
- **Parallelism:**
  - Performance difference is insignificant using AVX and AVX2 for NASA applications . AVX2 uses more power therefore core frequency reduces from base frequency to AVX frequency.
  - Hyper-Threading (HT) degrades the performance of NASA applications.
- **Turbo-Boost**
  - Turbo-boost is effective only for few cores. On one core performance by TB for compute intensive kernels is 3% to 8%. On most NASA applications It has no impact on NASA applications using all the cores of a node.
- **Modeling:**
  - It is hard to get even double digit floating-point efficiency on kernels like AXPY and DOT product (8%).

Questions: Send email to Subhash Saini at subhash.saini@nasa.gov